Liver Disease Prediction Using Bayesian Classification
S.Dhamodharan
Assistant Professor, Department of Computer Applications, Easwari Engineering College

Abstract: Data mining is an activity of extracting meaningful data from a large database, using any of its techniques. In this paper we use Bayesian classification technique, which is one of the major classification models. Unlike the other disease prediction, here we predict three major Liver Diseases such as Liver cancer, Cirrhosis and Hepatitis with the help of distinct symptoms. The primary goal is to predict the class type from classes such as ‘Liver Cancer’, ‘Cirrhosis’, ‘Hepatitis’ and ‘No Disease’.

General Terms: Data Mining, Liver disease, Classification.

Keywords: Naïve-Bayesian, FT tree, Weka tool.

1) INTRODUCTION

Data mining application includes various methodologies that have been developed and used by both research and commercial centers. These techniques have been used for banking, industrial and other scientific purposes. This research determined that whether data mining technique could also used for finding the type of Liver Disease using distinct symptoms. The Classification is based on supervised learning.

Liver is the largest gland in our body with the size of the football and about 1.36g in weight. The liver lies on the right side of the abdomen and beneath the diaphragm. The Liver’s main job is to filter the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs; it also secretes bile and makes proteins important for blood clotting and other functions.

The Liver is an important organ in our body, we can survive only one or two days if it shuts down. Fortunately, the liver can function even when up to 75% of it is diseased or removed. This is because of its amazing ability to create new liver tissue from healthy liver cells that still exists.

Major Cause of Liver Disease:

When liver becomes diseased it can cause serious damage to our health. There can be several things and health conditions that can unknowingly cause liver damage.

- **Alcohol:** Heavy Alcohol Consumption is the most common cause of liver damage. When you drink alcohol the liver gets diverted from its other functions and focuses mainly on converting alcohol into less toxic form.
- **Obesity:** People who are obese have excess amount of body fat which tends to accumulate around the liver cause fatty liver disease.
- **Diabetes:** Having diabetes increases the risk of liver disease by 50 percent. Increased level of taking insulin results in fatty liver disease.

Common Liver Disorder:

**Hepatitis:** A disease which is caused by a virus spread due to sewage contamination or direct contact with the infected bloody fluids.

**Cirrhosis:** It is the most serious liver disease occurs when normal liver cells are replaced by scar tissue as the chronic liver disease.

**Liver Cancer:** The risk of having liver cancer is higher for people who have cirrhosis and other type of hepatitis.
2) DATA DESCRIPTION
The Datasets were the distinct symptoms of liver diseases such as ‘Obesity’, ‘Yellow Eyes’, ‘Loss Of Appetite’,'Diabetes’, ‘Pain in Abdomen’, ‘Weight Loss’, ‘Enlarged Spleen’, ‘Enlarged Liver’, ‘Fever’. Along with symptoms the other attributes is ‘Age’, ‘Gender’, ‘FAC (Frequent Alcohol Consumption)’ totally 12 attributes with 54 instances.

The main goal is to create an user friendly web application for liver disease prediction .The real medical data are provided only after the consultation of the physician, which would not be useful for prediction.

3) CLASSIFICATION IN DATA MINING
Data Mining has two major functionalities. The functionalities are Classification and clustering techniques. Classification techniques are mainly designed for classifying unknown samples using information provided a set of known classified samples. This known classified sample is referred as training set as it used for training the classification technique how to perform its classification. This research used classification technique to produce the result. Algorithm is selected by evaluating each supervised machine learning which done using weka tool. The goal of classification is to predict the target class for each case in the data.

4) ALGORITHM SELECTION
The basic classification is based on supervised algorithm. The algorithm is applicable for the input data. Classification is done to know how the data is being classified; the Classify Tab shows the list of machine learning algorithms. These algorithms run multiple times manipulating input data to increase the accuracy of the classifier. Weka is included with learning performance evaluators. The data sets are split into training and testing datas.

A comparative study among Naïve Bayes, FT tree is done. The objective of this comparison is to find out the best accuracy value and the best efficient technique to do the Liver Disease prediction.

4.1) Naïve-Bayes
Naïve Bayes is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. Naïve Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. An important advantage of Naïve Bayes classifier is that it requires only small amount of training data set to estimate the parameters necessary for classification. The formula for Naïve Bayes classifier is: 

\[ P(H | E) = \frac{P(E | H) \times P(H)}{P(E)} \]

The basic idea of Bayes rule is that outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed. A priori probability is known as an event before the evidence is observed. A posterior probability is known as a probability of an event after the evidence is observed. The experiment of Naïve Bayes is carried out using WEKA (Waikato Environment for Knowledge and Analysis) , with full data set for training and testing purposes. The accuracy of the algorithm is found out using weka tool which is compared to the other algorithms, so that the efficient algorithm can be found and later it can be used for prediction.

4.2) FT Tree
The FP growth algorithm is another way to find frequent item sets without using candidate generations, thus improving performance. It uses divide and conquer strategy. The main idea of this method is the usage of a special data structure named frequent pattern tree, which retains the item set association information. The advantage of FP Tree is a highly compact representation of all relevant frequency information in the data set. A great advantage of FP Tree algorithm is that overlapping item sets share the same prefix path. As the main disadvantage of FP-Growth is that it is very difficult to implement because of its complex data structure.

5) EXPERIMENTAL RESULTS
The paper deals with Naïve Bayes and FT Tree algorithm. Experimental Process is discussed using 29 training data sets and the results are compared. The performance analysis is done by considering only the accuracy of the model. The data analysis processed using WEKA data mining tool for exploratory data analysis, machine learning and statistical learning algorithms. The training data set which is used for finding the accuracy consist of 29 data sets with 12 different attributes. The attributes used for training sets for Liver disease prediction. With the help of these attributes the prediction is done on Testing Process. The performances of these algorithms are evaluated and the results are analyzed. Initially, the data are pre-processed to make the data mining process more efficient. With the help of WEKA the accuracy of the algorithms are compared. The preprocessed data is then classified using WEKA tool. The algorithms can be applied directly to the data set. WEKA contains tools for data classification, clustering, associate and visualization. In this paper we concentrate on Naïve-Bayes, and FT Tree. Classification is done to know the exactly how data is being classified. The result obtained from the weka tool is compared to the real scenario and the
algorithm which considered to more accurate is considered as the best algorithm and chosen for predicting the test attributes. The below diagrams shows the accuracy of these algorithms from that comparison is done and result is obtained.

**Naïve-Bayes Algorithm**

This algorithm classified 23 correctly instances from 29 instances. Kappa is a measure of agreement normalized for chance agreement.

\[
P(A) - P(E) \\
K = \frac{1 - P(E)}{1 - P(E)}
\]

Where \(P(A)\) is the percentage agreement (e.g., between your classifier and ground truth) and \(P(E)\) is the chance agreement. \(K=1\) indicates perfect agreement, \(K=0\) indicates chance agreement. The value of kappa statistics is 0.67%. The mean absolute error is the sum over all the instances and their AbsErrPerInstance divided by the number of instances in the test set with an actual class label (that should normally be all of them). MeanAbsErr = Sum(AbsErrPerInstance) / # instances with class label. The value of mean absolute error is 0.16. The accuracy of the Naïve Bayes classification which is applied on the training set is found out to be 75.54%.

- **TP Rate**: rate of true positives (instances correctly classified as a given class)
- **FP Rate**: rate of false positives (instances falsely classified as a given class)
- **Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class
- **Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)
- **F-Measure**: A combined measure for precision and recall calculated as \(2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})\)

**FT Tree**

This algorithm classified 19 instances correctly from instances. The Kappa Statistics value is 0.0794 and the value of mean absolute error is 0.4655. The accuracy of this algorithm is found out to be 72.6624%.

### 6) CONCLUSION

Data mining plays a very important role in scientific purpose. With the help of data mining, people can know the possible outcome of the Liver disease prediction. These days, the exploitation of knowledge is really recognized widely. In this paper Naïve Bayes, and FT Tree algorithm accuracy are compared and the result is obtained. The result tells that the accuracy of Naïve Bayes algorithm is much better than the other algorithms. Based on the experimental results the classification accuracy is found to be better using Naïve Bayes algorithm compare to other algorithms. From the above result we can conclude that Naïve Bayes algorithm plays a key role in predicting the result very accurately.

### REFERENCES